Meaningful Explanation Effect on User’s Trust in an AI Medical System: Designing Explanations for Non-Expert Users

RETNO LARASATI, Knowledge Media Institute, The Open University, United Kingdom
ANNA DE LIDDO, Knowledge Media Institute, The Open University, United Kingdom
ENRICO MOTTA, Knowledge Media Institute, The Open University, United Kingdom

Whereas most research in AI system explanation for healthcare applications looks at developing algorithmic explanations targeted at AI experts or medical professionals, the question we raise is: How do we build meaningful explanations for laypeople? And how does a meaningful explanation affect user’s trust perceptions? Our research investigates how the key factors affecting human-AI trust change in the light of human expertise, and how to design explanations specifically targeted at non-experts. By means of a stage-based design method, we map the ways laypeople understand AI explanations in a User Explanation Model. We also map both medical professionals and AI experts’ practice in an Expert Explanation Model. A Target Explanation Model is then proposed, which represents how experts’ practice and layperson’s understanding can be combined to design meaningful explanations. Design guidelines for meaningful AI explanations are proposed, and a prototype of AI system explanation for non-expert users in a breast cancer scenario is presented and assessed on how it affect users’ trust perceptions.


Additional Key Words and Phrases: Explanation, Trust, Artificial Intelligence Explanation

ACM Reference Format:

1 INTRODUCTION

While the interpretability and explainability of AI has been a concern since the early age of AI [92, 96], it has recently acquired particular importance as AI technologies are being deployed in a much broader set of societal contexts. According to the US Defense Advanced Research Projects Agency (DARPA) [35], Explainable AI is essential for a machine learning system to “enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners” [35]. Explanation is claimed to be able to solve the Fairness, Accountability, Transparency, and Trust problems that are becoming more apparent as the progress of AI development in different fields continues [40, 62]. Moreover, law makers and regulators now legally recognise to all users the right to algorithmic explanation, as specified by the General Data Protection Regulation (GDPR) in Europe, Equal Credit Opportunity Act in US, Digital Republic Act in France, and Personal Information Protection Law in China.

Authors’ addresses: Retno Larasati, Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, United Kingdom, retno.larasati@open.ac.uk; Anna De Liddo, Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, United Kingdom, anna.deliddo@open.ac.uk; Enrico Motta, Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, United Kingdom, enrico.motta@open.ac.uk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.

Manuscript submitted to ACM
At present, in Articles 13-15 and 22, GDPR mandates that “meaningful information about the logic” and “the significance and the envisaged consequences of such processing” of automated systems be made available. It’s noteworthy that the GDPR does not offer specific guidelines regarding what constitutes ‘meaningful information’ or the extent of transparency required, leaving room for interpretation. Notably, GDPR primarily affects laypersons as data subjects. It is crucial to emphasise that the provision of clear and comprehensible explanations, particularly tailored for laypeople, holds significant importance in the context of GDPR compliance. The obligation to provide meaningful AI explanations aligns with the broader ethical imperative to ensure the responsible development and deployment of AI systems in decision-making processes. Any institution that uses an algorithm to assist in decision making is therefore not only required to provide a meaningful explanation, but is also ethically called to build trustworthy AI systems. An AI system should be trustworthy, as it can contribute to mitigating the risk of potential misuse, overuse, or underuse.

However, providing explanation in an AI system is not a simple matter. There is no one-size-fit-all explanation [8, 93] since explanation needs are diverse and personal. How people interpret explanation can differ from one to another depending on various factors, including expertise. Users with different professional and epistemic backgrounds might perceive explanation differently [28] and non-expert users might have different criteria on what kind of explanation they may expect and understand. Whereas some users may require specific types of justifications or empirical reasons to believe that the algorithm’s or computer system’s assertions are valid and true, other users may easily accept algorithmic outputs as true and valid, or may even prefer an algorithmic decision to human advice [56, 63]. Additionally, it is also one of the essential requirements for a medical software or app to consider the technical knowledge, experience, and education of intended users (lay or professional) when designing medical software or apps, according to the UK’s Regulatory Agency MHRA Guidelines[71]. Previous researches in explainable AI were claimed to be not strongly informed and involving real users [1], including in explainable AI for healthcare [4].

The lack of explainability, transparency, and human understanding of how AI works are key reasons why people have little trust in AI healthcare applications; however, research indicates that transparency [40] and understandability [62] can be effectively used as means to enhance trust in AI systems. Nonetheless, the relationship between trust and explanation is far from being unidirectional. Recent studies show instances in which explanation can hinder rather than improve trust. For instance, Lim and Dey [58] have shown that providing explanation can improve trust because of increased user understanding; however, it could decrease trust because of exposed incompetency. Research evidence also indicates that in some circumstances, users tend to over-trust and continue to rely on a system even when it malfunctions [24]. Hence, in our study, we asked to what extent explanations can be used as sense-making aids, consisting of key pieces of information which can facilitate users to rethink their perception of an AI system, reflect on their initial trust, and make considered trust judgments. We formulate the following research questions: How can we design meaningful explanations for AI healthcare non-expert users (i.e., patients or carers)? and What are the effects that meaningful explanations can have on users’ trust judgement and perception towards an AI system?

The specific aim of this paper is to navigate the complexities of designing explanations that are adequate and understandable, or what we call meaningful, for laypeople. Existing research, such as the work by Ehsan et al. [28], emphasize the importance of tailoring explanations to the specific needs of users. To do so we engaged with three types of stakeholders: domain experts, AI experts and laypeople. In fact, insights from AI/Machine Learning (ML) developers, medical professionals, and users are equally essential to the design of meaningful explanations. Medical professionals have knowledge of traditional healthcare scenarios and have established guidelines on how to explain a diagnosis or health condition to patients. AI experts have extensive knowledge of the AI system, its processes, and what technical aspects can be explained to laypeople. Non-expert users, on the other hand, are asked to understand an AI system’s
Meaningful Explanation Effect on User’s Trust in an AI Medical System

105 explanation with minimal (or zero) technical and medical knowledge. Hence, they are unique source of information
106 about issues, needs and expectations about the medical explanation process from the patient perspective [11, 49, 58, 82].
107
108 To design meaningful explanations involving stakeholders with a diverse expertise profile, we took a Human-
109 Computer Interaction (HCI) approach to explanation design [1, 35] and adapted a stage-based design process previously
110 proposed by Eiband et al. [29]. The method was successfully implemented for explanation design of a recommender
111 systems for fitness applications, and it is particularly suitable to our case since it enables individual investigation of
112 expert and non-expert views on the problem. It also provides a framework to combine different stakeholders’ knowledge
113 to inform design guidelines. The method consists of five stages and can be divided into three main phases. The first
114 phase focuses on “what” to explain, through the construction of an Expert Explanation Model (what “can” be explained)
115 and a User Explanation Model (what “needs” to be explained). It then focuses on synthesising the two models in a
116 Target Explanation Model, which can be translated to design guidelines that describe “how to convey the explanation”
117 (See Figure 1).
118
119 The Target Explanation Model is then used to inform the next phase, the development of explanation design
120 guidelines, from which we developed a prototype Explainable AI system for breast cancer assessment. We focused on a
121 self-managed breast cancer assessment scenario, in which the results of thermography scans are automatically analysed
122 by an AI system and need to be communicated and explained to the prospective patients. We developed a prototype
123 system which provided explanations that adhere to our proposed explanation design guidelines.
124
125 Finally, in order to assess both the prototype system and the explanation design guidelines from a layperson
126 perspective, we went back to the non-expert users we previously interviewed, showed them the prototype, and asked
127 feedback on their understanding of the system outputs and whether they felt it enabled them to make considered trust
128 judgements.
129
130 Our findings show that the explanation prototype designed following the proposed design guidelines for AI explana-
131 tion was considered meaningful by non-expert users. Furthermore, results suggest that our explainable AI prototype
132 system helped non expert users to make considered trust judgements toward the AI system recommendations, decreasing
133 the number of people over-trusting or distrusting the system.
134
135 The contribution of our research is threefold. Firstly, we propose three explanation models (expert, non-expert
136 and target explanation models) which conceptualise both AI explanation’s understanding and needs from expert and
137 non-experts users and how these can be effectively combined. Secondly, we propose a set of generic design guidelines
138 for meaningful explanation in healthcare, which we exemplified and assessed by developing a working prototype
139 explanation system for AI breast cancer self-assessment. Finally, from the evaluation of the prototype, we provide
140 an initial validation of the explanation design guidelines and discuss meaningful explanation effects on users’ trust
141 perception.

2 BACKGROUND

2.1 Meaningful Explanation

Explanation can be seen as an act or can be seen as a product, that can be categorised as good or bad. A good explanation
142 is an explanation that feels right because it offers a phenomenologically familiar sense of understanding [3]. In this
143 paper, instead of good explanation, we will use the term meaningful explanation, to stress our interest and focus on an
144 adequate and understandable explanation.
There is no single agreed-upon definition of a meaningful explanation. Guidotti et al. defined meaningful explanation as explanation that is faithful and interpretable [34]. Thirumuruganathan et al. defined meaningful explanation as explanation that is personalised based on users’ demographic [99]. GDPR Articles 13–15 provide rights to ‘meaningful information about the logic involved’ in automated decisions, without specifically defining what is ‘meaningful information’.

In this paper, we define a meaningful explanation as one that is acceptable and understandable to the user. A meaningful explanation is characterized by its usefulness and relevance to the user, providing information that is valuable and applicable to their specific context. Additionally, it should exhibit accuracy and completeness, ensuring that all necessary details and factors influencing the decision-making process are adequately conveyed. Furthermore, a meaningful explanation should be presented in a manner that is easy to comprehend, utilizing clear and accessible language or visual representations. These criteria align with previous studies in the field. For instance, Miller [73] describes explanation as a process that facilitates the understanding of facts related to a given context, while Lipton [61] defines explanation as information that helps users comprehend the decision-making process of an AI system. These sources collectively reinforce the importance of a meaningful explanation being useful, relevant, accurate, complete, and presented in an easily understandable manner.

It has been argued that transparency and post-hoc interpretability in Explainable Artificial Intelligence (XAI) should be regarded as encompassing forms of objectual understanding, while the functional approach to understanding AI models offers a more limited, purpose-driven understanding.

In cognitive psychology, explanation can be classified into different types: i. Causal explanation, which tells you what causes what, ii. Mechanical explanation, which tells you how a certain phenomenon comes about, and iii. Personal explanation which tells you what causes what in the context of personal reasons or beliefs [104]. Approaching these definitions from a explainable AI and AI reasoning angle, we could say that causal and mechanical explanation could be the same, because the causal explanation of an AI system is mechanical by definition. For instance, if we ask why the AI system give us a certain prediction, the answer will consist of the AI’s mechanical process to produce that prediction result. Personal explanation might also not be relevant, since all AI “personal” explanations in terms of what causes what in the context of the specific AI reasoning or beliefs (assumptions) is also a specific type of mechanical explanation which looks into the specifics of AI reasoning mechanism. Therefore, we will focus on causal explanation next.

Hilton proposed that causal explanation proceeds through the operation of counterfactual and contrastive criteria [38]. Lipton suggested that “to explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q” [60]. Miller quoted Lipton and argued that everyday explanations, or human explanations, are “sought in response to particular counterfactual cases. [...]people do not ask why event P happened, but rather why event P happened instead of some event Q” [73].

Causal explanation happens through several processes [38]. First, there is information collection: a person gathers the information available. Second a causal diagnosis takes place: a person tries to identify a connection between two events/instances based on the information. Third, a cause is selected, a person dignifies a set of conditions as ‘the explanation’. This selection process is influenced by the information gathered and the domain knowledge of a person [69]. This means that what people consider acceptable and understandable would be selected from the information provided and would be dependent on their own domain knowledge or role.

According to Lambrozo, explanations that are simpler are judged more likely to be believed and more valuable [64]. A study about simple and broader explanation also highlighted that users prefer a combination of simple and broad explanations [84]. As mentioned previously, explanation can be seen as an act or can be seen as a product. Explanation...
as an act involves the interaction between one or more explainer and explainee [73]. According to Hilton, explanation is understandable only when it involves explainer and explainee engaging in information exchange through dialogue, visual representation, or other communication modalities [38]. This statement implies that static explanations could be harder to understand because they could be less engaging and would not involve a dynamic interchange between explainer and explainee. To achieve meaningful explanation, a social (interactive) characteristic of explanation needs to be taken into account.

Previous research also showed that participants place the highest trust in explanations that are sound and complete [51]. Soundness here means nothing but the truth, how truthful each element in an explanation is with respect to the underlying system. Completeness here means the whole truth, the extent to which an explanation describes all of the underlying system. Completeness is argued to positively affect user understandability [50]. Moreover, in the context of AI explanation, Paez’s work [79] argues that the functional approach to comprehending AI algorithm, which when understandability focuses on the purpose of the algorithm and the relation between its features and decisions, offers a more limited understanding geared toward specific purposes. Although Kulesza’s studies of explanation apply to music recommender systems, we believe that being truthful (soundness) and thorough (completeness) are key characteristics of explanations to be further explored.

2.2 Why Explanation Design for Non-Experts

Explanation is a complex phenomenon and highly context and users’ dependent. In 1983, Clancey [23] carried out seminal research to examine explanation methods of a clinical decision support system called MYCIN. The system was capable to diagnose patients based on reported symptoms and medical test results, using 500 rules, and if requested, MYCIN could explain the reasoning that led to its diagnosis. MYCIN’s explanation design goals was to be educational to non-expert medical students. However, the students found MYCIN’s explanation to be insufficient. Nowadays, explanation systems do not necessarily aim to educate users, but rather to solve the black-box problem of AI recommendations and gain users’ trust. Even though the goals of our research are different and the target users are non-experts in general, important insights from Clancey’s research still stand today. In particular that explanation design is expertise dependent: non-expert users, medical doctors and medical students have different standard of explanation, which means the explanation design should be user specific and targeted with a more participatory approach.

Researchers have explored different user-centered approaches to design explanation for different types of AI systems [58, 59, 82], including AI healthcare systems [22, 43]. Despite many approaches have been proposed, explanation design specifically targeted to non-experts or laypeople has received scarce attention. Research on explanation design for AI healthcare systems mostly targets expert users [15, 20, 103]. This highlights a research gap in AI explanation design that can be meaningfully interpreted by laypeople, and effectively used to calibrate their trust toward the system. On the one hand, AI developers and designers cannot just apply an explanation framework designed for medical professionals and give it to non-experts expecting them to make sense of it. On the other hand, improving laypeople’s understanding of the AI systems could positively affect their degree of trust in these systems [66] and their recommendations [25]. The next section lays the grounds of our rationale for the importance of investigating the link between trust and explanation.

2.3 The Importance of Human Trust in AI

Users’ trust is one of the crucial factors influencing the adoption of AI systems. When it comes to AI systems in healthcare, trust is even more essential, since AI is applied in sensitive contexts such as helping to diagnose diseases, developing new drugs, and gaining better insights into treatments and prevention that can benefit society as a whole.
Trustworthiness, seen as a characteristic of a system perceived by a user as worthy of being trusted [44], is increasingly becoming an ethical and societal need, and it is even more so in healthcare where using AI adds an element of uncertainty and risk for the vulnerable patient [6]. Trust is humans’ primary reason for acceptance [31], without which the fair and accountable adoption of AI in healthcare may never realise. However, there is evidence in the literature that both medical professionals and patients show little trust in AI healthcare systems, which might hinder their broader adoption in society.

In particular, a 2018 survey study found that 30% of clinician respondents lacked trust in AI [41]. Such lack of trust also extends to famous AI systems from reputable companies such as IBM. Watson for Oncology was promoted by IBM to cancer doctors but many doctors from partnered hospitals chose to ignore the system recommendations. Watson provided supporting evidence for the recommendations it made, but did not actually explain how it came to recommend a particular treatment for a particular patient, which left doctors confused[1]. In addition, it is not only doctors who are skeptical, patients are also unsure about the use of AI in healthcare. In a 2016 survey, 61% of UK respondents declared not to be willing to engage with AI for their healthcare needs[2], while a 2019 survey reveals that only 20% of US adults would trust AI-generated advice for healthcare information[3]. There is also a factor of distrust that came from the healthcare system itself. Cases of AI system in healthcare were reported for discriminating against specific ethnicity group, religion or body mass index[75, 76], something that may have been caused by biased data imported from a real health care system[80]. Unfortunately, the specific reasons why individual patients or non-expert users are sceptical about trusting AI in healthcare can only be speculated.

In order to understand what does it take to trust an AI system in the healthcare domain, we need to understand in what circumstances it is appropriate for a human to trust AI results or recommendations. The trust between human and computer/machine has been widely studied and research shows that human trust is often dependent upon people understanding of the system [58, 74, 85]. People, in general, and medical professionals in particular, may resist adopting an AI healthcare system if they do not understand the system’s capabilities and intended use [68, 102]. To address this issue, researchers are working to find out what information users want to understand from AI systems and how to present it [15, 37, 82, 106]. However, there is also the issue of over-trusting, otherwise known as “automation bias” when people are overly relying on algorithms [24, 26].

2.4 The Relation between Explanation and Trust in AI systems

Researchers have tackled the issues of over-trusting, automation bias, distrust, and algorithm aversion; by the role of explanation in calibrating users’ trust, and in helping users to make a more informed decision on whether to trust or distrust the AI system [15, 100, 110]. Simple algorithm explanations, such as algorithm confidence level or algorithm accuracy percentage, have shown to affect users’ trust perception, in some cases increasing [53, 108, 109], in others decreasing [27, 58], yet in others not affecting users’ trust [110]. Some research argued that providing a more thorough algorithm explanation, such as detailed algorithm features could help user calibrate their trust judgement [16, 85]. However, only providing algorithm’s explanation might be insufficient for users, especially for AI in healthcare [4]. In this paper, we focus on designing an explanation outside of the algorithm explanation or XAI. We aim to investigate how to design a meaningful explanation, which is adequate and understandable for non-expert users, and see if meaningful explanations (of the output, not the algorithm) can be a tool to help users to better consider their trust judgements.

1https://www.statnews.com/2017/09/05/watson-ibm-cancer/
3https://www.invoca.com/blog/new-invoca-research-conducted-by-the-harris-poll

Manuscript submitted to ACM
Meaningful Explanation Effect on User’s Trust in an AI Medical System

towards an AI system. To achieve this we need to understand how non-expert and expert users understand and process explanation of AI healthcare systems and how explanations can be designed to respond to different users needs and expertise.

3 METHODS

To design a meaningful explanation for an AI medical support system, we carried out a stage-based design process adapted from Eiband et al. [29]. Eiband et al. successfully implemented a participatory design process which involved different stakeholders to design an explanation interface to improve system transparency from a user perspective [29]. This approach involves a design process to build an explainable user interface (UI) which aims to answer two key questions; “What” to explain and “How” to explain it. The “What” defines the content of the explanation and the “How” defines the presentation format of the explanation. In previous studies using the same method [29, 100], the explanation goal was system transparency, where transparency is the property of making the system inner functioning available and understandable to users. Transparency helps users build an internal conceptualization of the system, and it is important to bridge users’ and designers’ understanding of how the system works in terms of its internal structure and processes, or what we refer to as a mental model [45]. However, our explanation goal is not focused on transparency of the AI inner functioning (algorithmic explanation), but rather on providing adequate and understandable explanation of the AI system output (in this case an healthcare assessment). Our design process aims to build explanations that are not only focused on the detailed understanding and transparency of the AI algorithm [50], but more on eliciting user explanation requirements, in terms of: which elements of the system need to be explained, should be explained, and how they can be explained, would understandable from a layperson perspective. Thus, instead of developing mental models of the AI system, we built what we call Explanation Models consisting of required explanation components by both experts and non-experts.

To achieve this, we grouped and modified the five stages of Eiband et al. [29] method in three phases: 1. Explanation Models Development, 2. Design Guidelines and Prototyping, 3. Evaluation. This grouping better captures the research phases we followed, spanning from the development of the conceptual Explanation models (Stage 1-3), the translation of the Target Mental model in design guidelines and a working prototype (Stage 4), to the evaluation (Stage 5). In the Prototyping and Evaluation stages (4-5), we focused on translating the Target Explanation Model into a clearer set of design guidelines, that have been later implemented and tested. The final structure of our stage-based design process, the main guiding questions, and the data collection method for each stage can be seen in Figure 1. The detailed method for each phase and stage is explained in the next section.

MODELS OF MEANINGFUL EXPLANATION

4 EXPLANATION MODELS DEVELOPMENT

4.1 Expert Explanation Model

The Expert Explanation Model definition stage aims at capturing the experts’ understanding and vision of what an appropriate explanation of the results from an AI medical support system should look like in the case of laypeople. The experts involved in its development were both machine learning technologists and medical professionals. This research stage aimed at defining what can or should be explained to the wider public from an expert perspective, by distilling a series of explanation components, which represent the Expert Explanation Model.

Manuscript submitted to ACM
Data Collection and Analysis

Six participants were recruited by email, from the authors’ personal research and social networks; three were AI/machine learning practitioners with five years and more experience, and the other three were junior medical doctors in the UK and Indonesia. The main guiding questions that drove this stage were: what can be explained?; and what does an expert explanation for non-experts looks like? We asked the questions based on participants respective expertise (medical professionals and AI experts). We also showed participants two examples of breast cancer-related systems currently in commerce (Fig. 2), to understand how experts make sense of AI systems’ outputs and how they would explain the results to non-experts.

The interviews ran for 2 hours (max) each and were audio recorded. The recordings from the interviews were then transcribed. We used grounded theory approach with thematic analysis to analyse the interview transcripts, this involved developing codes, categories and themes inductively (open coding), rather than imposing predetermined classifications on the data [32]. Moreover, within healthcare domain, employing thematic analysis in the context of grounded theory provides a structured method for dissecting qualitative interview data [19].

Result

Three main coding groups emerged from the systematic qualitative data analysis: explanation content, which is specific about what information needs to be in the explanation; explanation customisation, which is the possible customisation that needs to be considered when giving an explanation; explanation interaction, which is possible interactivity that the user should have in the explanation.

We found that all medical experts gave very similar answers to the question of what can be explained to patients, which is in line with their medical training and their view of explanation as diagnosis delivery. In a common healthcare
scenario, before making a diagnosis, medical experts ask for patient symptoms and concerns and then confirm them back to the patients. When interacting with patients, the mandatory information they usually deliver to the patient are: disease information, possible treatments to choose, and the next steps for the patient to take. This is part of their medical diagnosis process in the information-gathering phase, including gathering test results. The information given by the patients can be interpreted as system input in the case of AI assisted self managed health, while the diagnosis system acts as the doctor.

How medical doctors deliver explanations to the patients is quite flexible and usually customised based on three different aspects. The first aspect is the patients’ requests and needs. Patients can request or ask for information and explanation of their specific needs. "It depends on how curious they are. If the patient just wants to know the diagnosis, then I may just tell them about it."- M3.

The next aspect is the diagnosis result. Since AI in healthcare is about people’s health, in potentially life-death situations, this aspect is highly sensitive. If the diagnostic result consists of a serious condition, reassuring words need to be included in the explanation to help patients feel less stressed and worried. On the other hand if the result is good, or if there is no sign of distress from the patient, there is less need for reassuring words. "I think one of the important things if it’s about serious conditions, we need to put more empathy."- M3. The diagnosis result can also be related to the patients’ request. When the diagnosis is for a serious condition, medical professionals give the patients’ options to choose what information they want to be disclosed. "if it’s like a common disease, which is easy to be treated, we usually give all of the information. but if it’s something serious, or if there’s no treatment for this, or serious suggestion, because not all people can accept this thing, again we ask them about how much they want to know about the information."- M1.

There is also an aspect of patients’ knowledge or education level perception. Medical experts mentioned that they customised the language they used and adjust the explanation complexity based on the their perceptions on patients’ knowledge or education level perception. They argued that some people who live in a rural area might have different knowledge than people who live in a big city, resulting in a simpler explanation. "People in the rural area, they don’t get the privilege to get a proper education, so it’s challenging for them to absorb the explanation. [...] So for example, “You have a lung infection, called tuberculosis. It is caused by bacterium blablabla.” for people in the rural area, they might not understand “doctor, what is an infection? What is bacteria?”- M2
After delivering information to the patients medical experts usually ask if there are any more questions. This interaction can happen back and forth several times until the patients have no more questions. “...Then we will explain what’s the next step. And we will ask if they have any questions or not. Including about the diagnosis and the plan.” - M3.

“We usually have direct face to face communication, so whenever patients ask, we then answer the questions directly.” - M1.

To summary, the key components of explanation, in terms of explanation content, required customisation, and possible interaction from medical experts perspective can be seen in Figure 3.

**EXPLANATION: Medical Expert Explanation Model**

<table>
<thead>
<tr>
<th>content</th>
<th>customisation</th>
<th>interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease information</td>
<td>Explanation/No Explanation : based on patient request</td>
<td>Input check</td>
</tr>
<tr>
<td>Disease treatment</td>
<td>Different explanation complexity: based on patient knowledge background</td>
<td>Open question</td>
</tr>
<tr>
<td>Next plan/step</td>
<td>Empathy (measuring words/ro: based on the result severity)</td>
<td>Detail request</td>
</tr>
</tbody>
</table>

Fig. 3. Expert Explanation Model: Medical Expert

Unlike in the case of medical experts, AI experts’ answers to the question “what can be explained” were quite diverse, and sometimes referred to their interpretation of medical knowledge. Their answers of explanation content for non-expert can be grouped into three components: system input, system process, and system output. “We have the inputs and intermediate results. The inputs are the different variables, as the driver for the predictor;” - A2.

The system input and output are quite obvious components to include, since users need to see what they want to analyse and the analysis results. Not as straightforward as input and output, the answers for system process were varied, from the specific algorithm explanations, such as AI features or calculation, to general explanation such as the algorithm accuracy or who made the algorithm. “if they’re trying to recognise cancer in a certain image, so this is the feature that helped me the most having this conclusion” - A1.

“You can try to show the formulation of the calculation. But some algorithm doesn’t provide how it works. If that is the case, you can just shift the responsibility to the one who made the algorithm; “this group of people made this algorithm.” - A3.

However, AI experts questioned the importance of providing algorithm explanation or the back-end logic to the non-expert users. In a real-life situation, AI experts claimed they rarely need to explain how the system works to a layperson unless they asked for it. “I have never met non-expert user that is interested in the artificial intelligence or the machine learning part of it, even the expert from the Ministry (people they work for), they are not really curious [...] not really interested or want to know about the algorithm, more about the accuracy.” - A2. “The context of working with user, if the app is working properly you don’t need to explain. But if there is a problem, you need to explain what is going wrong.” - A3.

AI experts also mentioned the option for users to request detailed information. It could be overwhelming for the user to read all the information in a comprehensive explanation, and it’s better to give the option for users to request
detail if they need it. "First we need the user to see the general output, but they can expand on some detail. Making it simple, just a few statements, and the general result, and if the user is curious, they can dig into it." - A3.

Input comparison was also mentioned as one of the components that should be included in the explanation. Visualization on how different inputs change the output could help the user to understand the system better. "the explanation is more about the user interface. Not only about the calculation, data, or number. They want to see the visualisation. If this then what." - A2. In summary, the key components of explanation from an AI experts perspective, in terms of explanation content, required customisation, and possible interaction can be seen in Figure 4.

Fig. 4. Expert Explanation Model: AI Expert

4.2 User Explanation Model

In the User Explanation Model definition stage, we captured users’ understanding and their perspective on how explanation should be presented in an AI medical support system. The purpose of this stage was to acquire knowledge about how do users currently make sense of explanations. This acquired knowledge was then structured in several key components of explanation, which constitute the User Explanation Model.

Data Collection and Analysis

To make sure that results of our interviews were not affected by pre-existent trust propensity toward AI, we sampled the participants based on their dispositional trust towards an AI medical support application and made sure there was a nearly equal number of people in each trust group (the AI sceptics, the open-minded, the AI enthusiasts). In fact, Szalma & Taylor (2011) showed that trust propensity is one of the human-related factors that could affect the response to an intelligent system [97].

To account for trust propensity, we recruited four participants each for three levels of dispositional trust, with 12 participants in total. Participants were recruited by email and social media, from the authors’ personal research and social networks who are not AI/ML experts and not healthcare professionals. To identify the level of trust, we asked the perspective participants to answer the following question: "if there was a cancer risk assessment/self-detection application available on the market, how likely would you be to use it? Please rate the likelihood from 1-7". This question
was sent in advance of the interview invitation. The participants were then grouped into three groups, the sceptic (1-3 likelihood responses), the open-minded (4-5 likelihood responses), and the enthusiast (6-7 likelihood responses). We sought to balance out the age range (twenties to forties) because research suggests that age can affect users’ trust towards a system, where older adults are more likely to trust the system than younger adults in a medical management system (decision aid) [39]. We also balanced out the male-female participants by recruiting one male in each group. We did this because we recognised that, despite male breast cancer only accounts for less than 1% of all breast cancer diagnoses worldwide [107], men may often act as carer and, as such, be included in the decisions taken by affected women close or related to them.

We followed the same interview structure as in the experts’ interviews. The main guiding questions we asked the non-expert users were: how do users currently understand AI explanations?; what does a user explanation looks like? We then showed the participants two examples of breast cancer-related systems to probe non-expert users reflections and feedback on the AI system’s results and explanations. The interviews were audio recorded and ran for 2 hours (max) each. The recordings from the interviews were then transcribed and analysed in a grounded theory approach using thematic analysis [19].

Result

From the systematic qualitative data analysis, seven groups emerged that can be classified under explanation content. Disease information, such as disease name, symptoms or severity level were the main information participants would like to receive from an explanation. Participants would also like to know about Disease Treatment options and what are the Next steps or actions that they could or should take. “you got cancer, and your options would be these, these, and these, and this is how I want to proceed. This is your options.” - E2. “After I got the result, do I need to contact my physician directly or is there a next step that is also provided by the app itself? Because it’s an entire journey, right?”- OM1. The use of graphics and images in an explanation was also suggested, as a way to help them understand the result, to show the comparison of contrasting condition. “Perhaps have some examples of how affected breast looks like, how unaffected breast looks like. So you can compare yourself with what is being put in your input.”- E2. “and then the image comparing, you know, both, my results and the healthy ones.”- S1.

Other than health related information, participants also want information related to the AI system to be included in the explanation. One of the information is System process. “I would want to know, what are they doing actually in the background to do this?”- E1. However, not all of the participants expressed interest in knowing the system process and some were not keen on knowing the background information. They argued that in a stressful situation, such as cancer positive, their focus would be less about the system information and more about their well-being. “Says that I have cancer, then I am not going to be interested in the system process”- S1. This sentiment supported previous AI expert opinion stating that non-experts are not interested in knowing the technical side of a system. The reluctance in knowing the technical information is not only a matter of timing, but it also reflects their reluctance caused by the possibility of not understanding the technical terms used to explain the process. “I mean, the very hard fine grain details? it will be incomprehensible for me because I am not familiar with the technology and everything.”- E2.

General System information, such as information about the Data used by the AI system, was also mentioned several times by the participants. The information can be about the data volume and data privacy. “I have to know how big their database.”- OM2. “And how am I sure that my breast picture will not be leaked to be utilised for other intentions and such.”- OM1. “Where are these data going?”- OM4. Next information is AI system Accuracy. “However, for my health, I think it
will be quite beneficial if I know how accurate it can be" - OM1. “Accuracy. Yeah, it’s better to put that. And if it’s a hundred per cent, I don’t think I will believe it. It’s better if it’s less than a hundred.” - OM2.

Other than the content of explanation, how the explanation is presented was also a concern for the participants. Participants expected care and empathy from the AI system, especially if their result is a serious condition, in the form of word choice and delicate delivery. “it might be quite direct and aggressive to say to the user, you have a cancer exclamation mark, or maybe you should keep it in according to these systems, it’s a high probability, that you might develop cancer in the next five years. Be a little bit more reserved, rather than explicit into your statements because it’s quite sensitive.” - E2.

From the interaction perspective, participants said that explanation should include an option to make a Doctor appointment. Participants also preferred to be able to request a more detailed information if they wished so rather than being presented with the full explanation in one go. The final User Explanation Model distilled from the interviews can be seen in Fig 5.

![EXPLANATION: User Explanation Model](image)

Fig. 5. User Explanation Model

### 4.3 Target Explanation Model

In the Target Explanation Model definition stage, we identified the key components of an explanation that users might want to include in the AI explanation User Interface (UI). The Target Explanation Model was built by including the highest common denominator between all components emerging from expert and non-expert understanding so to build an explanation model which takes full consideration of the views of all participants, in the basis of their different expertise (Medical Professionals, AI experts and Laypeople). The purpose of this stage is to ascertain and establish the focus of our explanation design, which represents both experts’ and layperson’s understanding and is targeted to non-expert users. Since our design goal is meaningful explanations, meaning adequate and understandable explanations, experts’ practice and laypeople’s considerations are important. The Target Explanation Model determines the focus of the explanation design by integrating Expert Explanation Model’s components with the components from User Explanation Model.
Data Collection and Analysis
We conducted semi-structured interviews with the same group of non-expert participants involved in the User Explanation Model definition, 12 participants with three levels of dispositional trust. The main guiding question was: which component of explanation should be realised in the User Interface to explain AI results? We showed the participants the combined version of Expert Explanation Model from medical and AI experts. We then asked participants to reflect on the explanation components from the Expert Explanation Models and discuss the components’ importance and value. Participants were asked to explicitly reflect on each explanation component by giving a rating of importance, 0-not important to 10-important, and express their reasoning. Based on the critical analysis of User and Expert Explanation Model, combined with the rating scores from non-expert users, we distilled the Target Explanation Model.

Result
The lowest rate given to a component by the participants was given to the need to customise the explanation to the users’ background knowledge (complexity customisation based on user education/knowledge). Participants deemed that explanation ought to be understandable for all users, regardless of their educational background. They also had diverging opinions on the need to request for explanation, information about system process, and the ability to ask open questions. As a result, these three components scored lower in term of importance. Indeed participants argued that explanation should be available whether the user requested it or not. Not all participants were interested in knowing the technicality of how the AI system makes decision/prediction, reflecting the same sentiment expressed by AI experts (See Expert Mental Model subsection). Some participants were also sceptical about openly asking questions to the AI system and preferred to wait to ask the doctor at their next appointment. All other components were scored essential by all non-expert users. The median value of ratings given by the users is reported in Table 1. In the Target Explanation Model figure, the components which did not fall in the high-rated category from the rating given by non-expert users are indicated with a lighter text (See Fig 6).

In the following section we show how we progressed from the development of the explanation models to prototyping and evaluation.

5 GUIDELINES DEVELOPMENT AND PROTOTYPING
By reflecting on the findings of the Target Explanation Model, we summarised a set of guidelines for designing trustworthy explanations for AI medical support systems. We define trustworthy explanation as meaningful explanations which enable more considered trust judgements toward an AI system. We then designed a user interface prototype based on such guidelines. We explored each guidelines’ presentation possibilities and the specific functionalities of the system that could realise them. We decided on a website where the user could carry out breast self-assessment based on screening images from their medical scan portable device. This hypothetical system was inspired by commercial
portable medical devices, such as Talos, Braster, and also promising research [67], which investigates new medical devices allowing individuals to do breast self-examination. The practice of breast self-examination empowers women to take responsibility for their health and help raising awareness among women at risk, as recommended by the World Health Organization (WHO). We chose to design a system using thermogram image scan, because portable devices or smart phones can be used to take thermogram image scans [67]. In recent years, studies on breast cancer detection using thermogram images have been conducted claiming very high accuracy (above 90%) [111] [30] [9], in some cases even as high as 100% accuracy [98]. The thermal images we used were obtained from an anonymous dataset made openly available by the Thermography Center of Memphis and widely used on the internet.

5.1 Guidelines Development

From the Target Explanation Model we inferred, what explanation contents should be included, how explanation should be delivered, and what interaction is needed while explaining AI results to non-expert users. The Target Explanation model captures both experts’ and non-experts’ views of non-experts users’ needs. From this, we propose 14 explanation design guidelines for AI medical support system explanation interfaces (See Table 2). The guidelines are grouped into three categories, that mirror the Explanation Models’ components sets: Content, Delivery and Interaction. Explanation Content is what should be included in the explanation. Explanation Delivery is how explanation should be communicated. Explanation Interaction includes additional actions to enhance the explanation quality [38]. In the following we describe and discuss each of the guidelines.

EDGI: Disease Information

General information about the disease, which includes the name of the disease, what it is, the symptoms and causes should be provided. Both experts and non-experts expressed the importance of providing patients with this general information. Non-experts said that they would like to know: disease name, symptoms and the severity of the disease. Even though not all AI medical support systems aim at pre-diagnosing, disease information is still a vital part of AI in healthcare and would still be relevant for other functionality, such as symptoms checker or health advice applications.


https://www.braster.eu/en/system-braster/what-is-braster

https://www.memphisthermography.com/breast-health.html
<table>
<thead>
<tr>
<th>Explanation Design Guidelines (EDG)</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDG1: Disease Information</td>
<td>General disease information e.g.: name, symptoms, caused</td>
</tr>
<tr>
<td>EDG2: Disease Treatment</td>
<td>Treatment options and information</td>
</tr>
<tr>
<td>EDG3: Next Plan/Step</td>
<td>Next step user could take following the result</td>
</tr>
<tr>
<td>EDG4: System Information</td>
<td>General system information e.g.: data, system accuracy</td>
</tr>
<tr>
<td>EDG5: System Input</td>
<td>The data user inputted to the system</td>
</tr>
<tr>
<td>EDG6: System Process</td>
<td>System algorithm or the technical process to get its results</td>
</tr>
<tr>
<td>EDG7: System Output</td>
<td>System result e.g.: pre-diagnosis, analysis, recommendation</td>
</tr>
<tr>
<td>EDG8: Empathy (Reassuring Words)</td>
<td>Delicately deliver the results with carefully selected words</td>
</tr>
<tr>
<td>EDG9: Simple and General</td>
<td>Uncomplicated wording that is acceptable for laypeople from various education background and level</td>
</tr>
<tr>
<td>EDG10: Input Check</td>
<td>For the user to check the input (is it correct or not)</td>
</tr>
<tr>
<td>EDG11: Doctor Appointment</td>
<td>For the user to make a doctor appointment</td>
</tr>
<tr>
<td>EDG12: Open Question</td>
<td>For the user to ask open questions</td>
</tr>
<tr>
<td>EDG13: Input Comparison (Visualisation)</td>
<td>For the user to compare the result with other data</td>
</tr>
<tr>
<td>EDG14: Detail Request</td>
<td>For the user to request detailed information</td>
</tr>
</tbody>
</table>

Table 2. Our 14 Explanation Design Guidelines for AI in Healthcare, categorised by information included, information delivery, and interaction included.

EDG2: Disease Treatment
The expansion should include information about treatment options that the patient can choose from. Medical experts claimed that the explanations they usually give to patients while delivering a diagnosis also includes possible treatments for the patient to choose from. Non-experts commented that they would like to know about the treatments that they should undergo after diagnosis, what options they have, and if they have to make an appointment with their doctor or physician. "you got cancer, and your options would be these, these, and these, and this is how I want to proceed. These are your options." - E2.

EDG3: Next Plan/Step
Information on next Plan/Steps that the patient should follow after receiving the AI results should be clearly communicated. Doctors confirmed that this is also a key step while delivering explanation in real-life healthcare contexts. Non-experts said they would like to know what they are expected to do after receiving their results, and what the plan is. This item is different from Disease Treatment, because the next plan/step could be additional to treatment and could include for instance contacting your doctor, book another test, or just maintaining a healthy lifestyle if treatment is not required. "do I need to contact my physician directly or is there a next step that is also provided by the application itself?" - OM1

EDG4: System Information
Explanation should include general information on the AI system. This could include system specifications, such as what data is used, how accurate the system is, what certification the system has, or who’s behind the development of the system. The need to communicate system information was mentioned by non-experts during the User Explanation Model interviews as something that could play a very important role in shaping their trust judgement. "at least I have to know how big their database is, you know, sort of make sure it is trusted" - OM3. "...who’s behind it? That’s one thing that I never see in explanations, you know, like really, like, who are the people behind it? Why did they make this thing? Are they, are they doctors? I do want medical technology that is created with the help of doctors" - E2. "When you open the app or whatever, there’s like an information about the credibility of the software. I think that’s very important. [...] saying that it’s..."
Meaningful Explanation Effect on User’s Trust in an AI Medical System

credible and proficiency testing and certification.”- OM2. “However, for my health, I think it will be quite beneficial if I know how accurate it can be.”- OM1.

Interestingly, system information did not come up during the interviews with medical experts. We asked a follow-up question to them about this. They claimed that they did not mention system information as part of patients’ explanation because for them to use a healthcare application, the application must be recommended by their healthcare authority (e.g., NHS in the UK). Hence, the system accuracy and credentials are somewhat implicit in the healthcare authority’s recommendation. Of course in the emerging healthcare market, where new products and services are continuously emerging also beyond institutional healthcare authorities, this information becomes increasingly relevant, for both patients and doctors, to make considered trust judgements on AI outputs.

EDG5: System Input

The data used by the AI system to generate the output is also a key information to provide in the explanation. System Input can be an image, text, or tables, depending on the data used by the AI system. According to the AI experts, many Explainable AI algorithms (i.e. LIME [85], Anchors [86]), already include input in the explanation to describe the process from input to output. Similarly transparent algorithms or frameworks can provide a very useful output which can directly contribute to this explanation component.

EDG6: System Process

System Process is the information about the system algorithm or the system’s technical process that produces its result. AI experts expect System Process to be included in the explanation. “...for example if they’re trying to recognise cancer in a certain image, this is the feature that helped me (the AI) the most having this conclusion.”- E1. Nonetheless, some experts and non-experts also expressed concerns about the need and interest of non-experts to know the technicality of how the AI system actually makes a decision/prediction. “I already got a diagnosis, how I feel... in that moment, I would be scared, I would say all other information is important to me except for the system process.”- S4. “I have never met a common user that was interested in the AI or the machine learning (part) of it. [...] they are not really curious.” - A2. In addition, this information is now increasingly required by law and should therefore be included in the explanation. The European Union’s General Data Protection Regulation (GDPR) requires AI systems to provide “meaningful information about their logic” and the European Commission Checklist for Trustworthy Artificial Intelligence (ALTAI) advises that explanation should always be provided to any user when AI is involved.

EDG7: System Output

The output of the AI system is perhaps the most crucial information delivered in the explanation, and it is the focus of the explanation itself. An AI System Output can consist for instance of a recommendation from a AI recommended system, or a pre-diagnosis from virtual clinician applications. Both experts and non-experts implied that they expect to somehow “see” the output/result when the explanation is delivered “Rather than I explain it, it’s the tool that explains it. What we are thinking, we need the user to see the general output”- A2. An important part of the explanation design therefore needs to focus on where and how the output can be visualised in the explanation interface.

EDG8: Empathy (Reassuring Words)

A key required component of the explanation delivery, and perhaps the hardest to realise, is the use of empathy and reassuring words while communicating the AI outputs to patients. Empathy is a key component of the explanation delivery in healthcare, which is also explicitly included in the Guide for Health Care Professionals [14]. Previous research on medical explanations [14] has analysed in details the aspects that doctors should consider when delivering

---


Manuscript submitted to ACM
a bad diagnosis. In line with this research medical experts confirmed that delivering diagnosis works differently if the
diagnosis is bad. "I think one of the important things if it’s about serious conditions, we need to put more empathy." - M3 The
explanation needs to be given in incremental stages, and many aspects, such as, attention to the environment, patient
state of mind and perceptions need to be considered. Doctors follow protocols which include empathetic understanding
and communication such as: establishing an appropriate space and being sensitive to patient needs, asking what the
patient knows and how much the patient wants to know; assessing how the patient processed the information, asking
if there are any questions, and finally acknowledging the patient’s response. Protocols with similar steps have been
proposed and tested in the literature [10, 83]. Non-experts also expect empathy, and ask for sensitive word choices to
deliver the diagnosis results delicately. "Be a little bit more reserved, rather than explicit into your statements because
it’s quite sensitive." - E2. One of the key challenges for designing trustworthy AI explanation is therefore delivering
empathetic explanations.

EDG9: Simple and General
Explanation needs to be delivered with simple and general wording, that is acceptable for laypeople from various
educational background and levels. It is important to note that explanation that is delivered with simple wording. This
requirement came directly from the medical experts, and align with research findings suggesting that explanations that
are simpler are also judged more likely to be believed and more valuable [64].

EDG10: Input Check
In the expert interviews the medical experts mentioned how they usually ask for confirmation about the patient data,
symptoms and worries before making a diagnosis. This is part of their doagnostic process in the information-gathering
phase, including gathering test results. This means that the input data used by the AI system is, in normal circumstances,
checked before diagnosis. In the context of an AI assisted process, it would be therefore important for users to be able
to check the input before or after any AI process.

EDG11: Doctor Appointment
The need for users to make a doctor appointment has been considered crucial by non-experts for the explanation to
be trustworthy. They expressed the need for a specific interaction to be included in the explanation interface which enables them to talk to a physician. "I have a first layer, which is knowing the diagnosis. and then second layer is to get my diagnosis through the physician[...] And I do not need to repeat whatever I’m saying again, because they have it on my apps." -OM1. Medical doctors also expected this interaction to be a given feature.

EDG12: Open Question
Open Question is the availability for users to ask open questions. Medical professionals highlighted the importance to
allow non-expert-users to ask open questions. They explained that after giving patients/users information, they always
ask if there are any more questions. This interaction can happen back and forth several times until the patients/users
have no more questions to ask. "...Then we will explain what’s the next step. And we will ask if they have any questions or not. Including about the diagnosis and the plan." - M3. Non-expert participants expressed their preference with talking to an AI system via chat-bot "you know, many times people feel more comfortable to talk to a chat-bot than a person because they don’t feel they’re judged" -OM4. Alternatively, they want to talk to the doctor via chat-box "if you want to know more and you want to explore more about your risk, in the app, it can be like that automated chat thing, they can directly connected to a health care professional who provides information like through the chat-box" -E2. However, this interaction requirement was rated relatively low by non-experts during the Target Explanation Model interviews. Some participants would prefer to wait to ask questions face-to-face to their doctor during their next appointment and some
were sceptical about asking open questions to an AI system.

Manuscript submitted to ACM
EDG13: Input Comparison (Visualisation)

Both experts and non-experts suggested the importance for non-expert users to be able to compare their results. AI experts commented that simulation features could be beneficial for non-experts to get a grasp of how the system works, which in turn can lead to build trust in the system. They suggested interactive solutions in which users can input different data to see the difference in results. Likewise, non-experts asked to see the opposite diagnosis case, to decide themselves if the result makes sense to them or not. "Perhaps have some examples of how affected breast looks like, how unaffected breast looks like. So you can compare yourself with what is being put in your input." - E2. "and then the image comparing, you know, both, my results and the healthy ones." - S1. Explanations should include interactions which enable users to see the contrast and form their own understanding of how the AI system works.

EDG14: Detail Request

Both experts and non-experts agreed that they wanted explanation that could expand on demand. As mentioned in the EDG12: Open Question, medical experts base their explanation on patient’s request. AI explanation interfaces should therefore enable users to request more information at several stages of the interaction. "It based on the patient’s question and response. If they ask, then I will tell them. But if they don’t ask anything, then I won’t tell them anything." - M2. AI experts also claimed that general explanation is good as long as users are able to read more about it. "general and truthful information is good, because it’s not long, because if it’s long they don’t want to read it. It’s the reason why a lot of news portal now separated one article to several pages, so it’d be easier to digest." - A2. One of the non-experts encapsulated the general idea of detail request, "I think it’s better to have general explanation or as people can understand, but just the details in other place. you know, some people are curious, some people are not." - OM3

Fig. 7. Care’s User Interface Prototype: Explanation Page - Top Part
5.2 Guidelines Implementation: Prototyping

The prototype was developed after six cycles of feedback between researchers involved in this study. Each cycle included the analysis of the proposed UI prototype, gathering feedback and proposing improvements in short focused discussions with the design team. The final prototype was informed by all 14 items in the design guidelines. We looked at content, presentation, colours, item placement, wording, and user interaction. We also took into consideration good practice in medical explanation research [10, 14, 83]. The final prototype described below consists of three web pages: Login page, Home/Profile page, and the main focus: Explanation page. In the Home page, user can click to choose the image to be analysed, which will take them to the Explanation page.

Fig. 8. Care’s User Interface Prototype: Explanation Page - Bottom Part

The top part of the Explanation page can be seen in Figure 7. At the top, we find information about the system’s accuracy and a notice on the possibility of the result being incorrect "Please note: CARE is only 98 percent accurate. Therefore, there is a 2 percent chance that our assessment is incorrect." This information covers EDG4: System Information and EDG8: Empathy (Reassuring Words). It provides users with a reminder that the AI output is not absolute, even before the user reads the result.

Next to the input data (the diagnostic image) (EDG5) we find the analysis result (EDG7). “Based on the data and the image, our system detects abnormality. It is highly likely that you have Ductal Carcinoma InSitu (DCIS).” In delivering
the result (EDG7) we chose specific wording, such as “highly likely” to reiterate the pre-diagnostic nature and level of uncertainty of the output. This, as one of the non-experts mentioned in the interview, is important to: “soften the blow” and reduce distress and anxiety that users may experience while reading the output.

Just below the output we find EDG1: Disease Information. The words we chose followed an empathetic approach, combined with the use of simple and general terms (EDG8 and EDG9). In the cases in which medical terminologies had to be included, we looked at reducing the number of technical terms and provided links to outside sources available to learn more about the term, at each user’s time and convenience. We carefully wrote the explanation to avoid additional stress to users, by adding reassuring facts “Please note, that DCIS is considered non-invasive or pre-invasive breast cancer”. Outside sources were also provided, to easily reach more detailed disease information (EDG14), such as links to Cancer Society and Cancer Research websites.

An input comparison visualisation (EDG13) follows the disease information. “How my image assessment compare to other cases”. In this visualisation, users can see a comparison between their image assessment and, for example, a normal breast image, or a more severe case. We included a detail request option (EDG14), to enable users to see more cases and comparisons. Treatment options (EDG2) are then provided below, including detail requests for patients to read more and better inform their treatments choices.

The rest of the Explanation page can be seen in Figure 8. After the treatment options are communicated, users can see what can they do next (EDG3: Next Plan/Step). In this case, the only action a user can take is contacting their doctor, which is related to the next function, EDG11: Doctor Appointment. The button we put there is “Share this assessment result with your doctor”, which allows to open a direct line with a doctor, rather than making an appointment.

A section to receive “Further Support” was added to the prototype to offer a more empathetic delivery of the explanation (EDG8). That part concludes the local explanation, that is the explanation directly tailored to the user and her/his personal AI analysis. A more global explanation is then presented as additional information. This includes EDG6: System Process and EDG4: System Information. Since this information does not directly relate to specific results, some participants might not be curious about it. Therefore, additional information was displayed by using an Accordion View graphic control element, which can expand or collapse, to reveal the explanation on demand. We put data related information after the system process description because both relate to system specification.

A Chat-box was finally put at the end of the page to address EDG12: Open Question. We chose a chat-box solution for the prototype because, as mentioned in Section 5 (EDG12: Open Question), users mentioned chat-box and chat-bot as desirable solutions. The prototype includes a “CARE Agent” (as AI chat-bot) and “Dr. Laura Smith” (as doctor). On the footer, we then find additional EDG4: System Information “Recommended by NHS”. This responds to experts and non-experts need for certification and credential of the AI output. Lastly, an EDG10: Input Check was put on the Home/Profile page as a pop-up window to make sure the user has clicked on the correct image they want analysed before they are redirected to the Explanation Page (See Figure 9).

GUIDELINES AND IMPLEMENTATION EVALUATION

Fig. 9. Care’s User Interface Prototype: Pop Up
6 GUIDELINES EVALUATION: HOW THE EXPLANATION DESIGN GUIDELINES BEEN EXHIBITED IN THE PROTOTYPE

The evaluation phase covered both the explanation prototype and the explanation design guidelines and consisted of two combined methods, an online survey and semi-structured interviews. The main focus investigated in this stage is: the quality of the explanation design guidelines and their implementation. To examine the quality of our proposed explanation design guidelines, we checked if each guidelines implementation was considered relevant in the explanation prototype. Additionally, we also investigated the relation between meaningful explanation and trust, which will be discussed in the next section.

6.1 Method

We developed an Online Survey to evaluate the prototype, and whether or not our explanation design guideline (EDG) components (See Table 2), are relevant for the user to understand and trust the AI system. The first survey structure is shown in Fig 10. We requested participants to evaluate the degree of relevance each of the explanation content component (EDG1-EDG7) held for their understanding and trust in the AI system, employing a 5-point Likert scale (1-not at all relevant, 5-very relevant). Additionally, we posed a different query for explanation interaction (EDG10-14) components. In this case, participants were asked to assess the extent to which each key component in the explanation result would improve their understanding and trust in the AI system, using a 5-point Likert scale (1-very poor improvement, 5-very good improvement). However, since the visual comparison (EDG13) was presented as content instead of interaction, we included the component to the first question. In total, we asked 12 questions, eight questions for EDG1-EDG7 and EDG13, followed by four questions for EDG10-14.

Fig. 10. First Survey Structure
The semi-structured interview was developed to further investigate if the explanation design guidelines informed by the target explanation model were appropriately exhibited in the prototype, and how the user perceived both the prototype and the explanation. The first part of the interview was about their initial opinion on the prototype and explanation, how it was presented, and if there were any improvement needs they could think of. In the second part of the interview, we asked participants to look at the explanation design guidelines and assess if all guidelines were addressed by the prototype and to what extent. The final part of the interview was to review and eventually improve the explanation design guidelines.

6.2 Data Collection and Analysis

For the online survey, participants were recruited on Mechanical Turk, with a survey set up using Google Form. We added check-in question in the survey to maximise participation quality and avoid bot submission by checking if the participant read the text in the prototype carefully (See Fig 10). Among the 82 subjects who took part, 22 failed to answer the check-in question, and there were also six duplicate entries. In total, 55 participants’ answers were analysed using descriptive statistics, by calculating each item’s mean and median.

For the interview, we contacted the same participants we interviewed in the explanation model definition stages, since they had an already deep understanding of the explanation requirements. In total, we interviewed seven participants; two medical experts, two AI experts, one sceptic user, one open-minded user, and one enthusiastic user. Participants were encouraged to interact with the Care prototype and then were asked for their opinion on the prototype.

The interviews were audio recorded and ran for 1-2 hours each. We transcribed and analysed the semi-structured interviews with a deductive approach. As mentioned in the previous section, in this stage we wanted to evaluate if the explanation design guidelines were appropriately exhibited in the prototype and also how the user perceived both the prototype and the explanation. We chose a deductive approach to code the interview data because we already had clear interpretative lenses (the Explanation Design Guidelines) that we wanted to look at to identify themes of interest [12].

6.3 Results

6.3.1 Online Survey. In this section, we present the results obtained from the online survey conducted to evaluate the prototype and assess the relevance of the Explanation Design Guideline (EDG) components in affecting users’ understanding and trust in the AI system. Table 3 summarises the median, mean, and standard deviation (stddev) of the ratings for each EDG component, specifically Explanation Content and the rating distribution is shown in Fig 11. The median ratings for all the explanation content components (EDG1-EDG7 and EDG13) ranged from 4 to 5, indicating a high level of perceived relevance by the participants. The highest median rating was observed for EDG1,2,3,6,7 (median = 5), followed closely by EDG4, EDG5, EDG13, all with a median rating of 4.

The mean ratings for these components ranged from 4.11 to 4.67, further supporting their overall relevance. The highest mean rating was obtained for EDG7: System Output (mean = 4.67), while the lowest mean rating was observed for EDG4: System Information (mean = 4.11). These results can be explained as participants/users naturally prioritise the AI system’s output when evaluating their understanding and trust in the AI system. The system output provides tangible results or outcomes that users can directly assess and comprehend. Therefore, it is reasonable for participants to perceive the system output (EDG7) as highly relevant in enhancing their understanding and trust.

On the other hand, the lower mean rating for EDG4: System Information (mean = 4.11) suggests that not all participants prioritise or find the technical details of the AI system to be crucial for their understanding or trust.
participants may focus more on the practical outcomes or utility of the system rather than the intricate technical information such as data. This variation in preferences and priorities among participants can explain the relatively lower rating for system information (EDG4) compared to other components.

Table 3. Median, Mean, and Standard Deviation from Explanation Components (EDG1-7 and EDG13) ratings.

<table>
<thead>
<tr>
<th></th>
<th>EDG1: Disease Information</th>
<th>EDG2: Disease Treatment</th>
<th>EDG3: Next plan</th>
<th>EDG4: System Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>mean</td>
<td>4.65</td>
<td>4.44</td>
<td>4.55</td>
<td>4.11</td>
</tr>
<tr>
<td>stddev</td>
<td>0.58</td>
<td>0.94</td>
<td>0.66</td>
<td>0.79</td>
</tr>
<tr>
<td>median</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>mean</td>
<td>4.22</td>
<td>4.38</td>
<td>4.67</td>
<td>4.38</td>
</tr>
<tr>
<td>stddev</td>
<td>0.92</td>
<td>0.83</td>
<td>0.51</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Table 4 summarises the median, mean, and standard deviation (stddev) of the ratings for each EDG component, specifically Explanation Interaction and the rating distribution is shown in Fig 12. The mean ratings ranged from 3.84 to 4.31. EDG12: Open Question had the highest mean rating of 4.31, suggesting that participants considered asking
question is a component that have a positive impact on their understanding and trust in the AI system. Similarly, EDG11: Doctor Appointment (mean = 4.07) and EDG14: Detail Request (mean = 4.22) received relatively high mean ratings, indicating participants’ perception on the effectiveness of the two components in enhancing users’ understanding and trust. EDG10: Input Check had a mean rating of 3.84, suggesting a slightly lower perceived improvement potential compared to the other components.

Overall, the results of the survey indicate that EDG11, EDG10, EDG14, and EDG12 were perceived as relevant components for improving users’ understanding and trust in the AI system. These components received consistent median ratings of 4 (good improvement), suggesting their importance. The mean ratings further support their positive impact, with EDG12: Open Question being rated the highest.

The results of the online survey indicate that all Explanation Design Guideline (EDG) components were perceived as relevant or important for improving users’ understanding and trust in the AI system. With consistent median ratings between 4 and 5 across these components, participants acknowledged their significance in facilitating comprehension and fostering trust. The relatively high mean ratings further underscored their positive impact. It is important for designers and developers to consider these guidelines to ensure effective explanations that meet users’ expectations and requirements. However, further research and refinement may be needed to explore the specific aspects of each EDG component.

<table>
<thead>
<tr>
<th>EDG11: Doctor Appointment</th>
<th>EDG10: Input Check</th>
<th>EDG14: Detail Request</th>
<th>EDG12: Open Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>mean</td>
<td>4.07</td>
<td>3.84</td>
<td>4.22</td>
</tr>
<tr>
<td>stddev</td>
<td>1.00</td>
<td>1.05</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 4. Median, Mean, and Standard Deviation from Explanation Interaction (EDG10, 11, 12 and EDG14) ratings.

6.3.2 Semi-structured Interview. In the first part of the interviews, we asked for participants’ initial opinions on the prototype and explanations. Overall, the prototype was received positively. All of the participants mentioned that the prototype was easy to use, and some mentioned that it was intuitive, "It is easy to use. It is intuitive. It’s pretty straightforward." - PA1, "as of now, it’s very concise, it’s very intuitive and the clear what it is" - PA2. Participants commented positively on the specific part of the explanation, for example, Disease Information (EDG1) "it’s good to have this general
information that is available after you received a diagnosis." -PM2, and Detail Request (EDG14) "I like the additional information is in the form of Q&A, and it’s a link if someone wants to expand on that." -PA2. Participants also mentioned that they could easily understand the explanation, with one of them specifically enjoying the Image Comparison (EDG13) feature "I like the comparison between the two because if you just look at your image alone and you are not in the medical field and don’t have medical information, you will not understand what the cancerous part in that image is. But based on the explanation here, you understand that cancerous breast is hotter and normal breast is colder." -PA1.

Participants also pointed out potential improvements to be made. Even if participant PA1 positively commented on the EDG13, other participants expressed opposite feelings. "what does thermogram mean? what does it mean that the cells are red?"- PS, "I don’t understand what the colours mean. I need more information. I know it shows the comparison, I need to know more. How you define hot?" - PE. More attention needs to be paid to how the comparison is presented.

Another improvement mentioned was the presentation of Disease Treatment (EDG2), specifically bullet points to present the treatment options. "The information is there, but it’s better to use bullet points." - S1, "I could pay more attention with points like bullet points or numbers. [...] and easier to read"- E1. On the basis of this feedback, the future prototype will include more information about how to read the image and also will present the treatment options in a more structured way.

In the second part of the interview, we asked participants to look at the explanation design guidelines and assessed if all guidelines were addressed by the prototype. From the transcript, we grouped participants answers into "present", "rather present", and "not present". We found that most of the participants’ answers indicated "present" (73 times) or "rather present" (20 times) on all guideline items, with the exception for Doctor Appointment (EDG11), which was indicated as "not present" (1 time). We speculate that this might be caused by how we presented the Doctor Appointment as "Share analysis result to your doctor". This label indeed can be misleading and lead to think that this function is only for sharing data and not to open a conversation with a doctor. Items that were rated as "rather present" provided suggestion for improvements to some of the guidelines. For example, a participant needed more information on the Disease Information (EDG1), "there was a name, but I don’t think there is a cause here. I think it would be nice to have." -S1.

The final part of the interview was to review the Explanation Design Guidelines. Most of the participants agreed that the explanation design guidelines are good and complete, while still need improvements in some parts. One of the improvements mentioned was the design guidelines’ item name, such as Doctor Appointment (EDG11) and also Detail Request (EDG14). As mentioned previously, one of the participants stated that they could not see Doctor Appointment because it was not presented there. Since we want to make this guideline as general as possible, and not all AI in healthcare would be able to provide a doctor appointment function directly, we might need to rename the guideline to "Channel to Doctor/Clinic". Some of the participants were pausing for a second before answering to the detail request question, and there was a possibility that they were confused with the naming. "Yes, but I am confused with the word "request"; I thought it would require the user to submit something."- M2. A better term that could be used for this guideline is ‘details-on-demand’, also suggested by Shneiderman’s visualisation mantra [90], which might be more commonly used and understood.

7 IMPLEMENTATION EVALUATION: HOW THE USER RECEIVED THE EXPLANATION

In this section, we investigate the relationship between meaningful explanation and trust in the context of our study. Our evaluation aims to address two key aspects: firstly, to assess whether the explanation we designed effectively conveys meaningful information to the users, and secondly, to examine the impact of such a meaningful explanation on users’ trust judgments and determine the direction of this effect. By exploring these dimensions, we aim to gain
insights into the effectiveness of our explanation design guidelines in affecting user trust and providing a basis for informed trust judgement in the realm of AI systems.

7.1 Method

7.1.1 First Survey. As mentioned previously, the first survey conducted in this study was a part of the online survey for evaluation of Explanation Design Guideline (EDG) components. In Fig 10, it is shown that in addition to the EDG evaluation, this survey included additional questions to investigate the relationship between meaningful explanation and users’ trust in the AI system.

To check if our designed explanation can be considered as meaningful explanation, we asked participants to rate the explanation based on the characteristics of meaningful explanation [54]. Participants rated their agreement if the explanation given by the prototype addressed the six characteristics of meaningful explanation, as in being Contrasting (describing something relative/in contrast to some other things), Domain-dependent (information relative to a specific background context), General (simple and broad), Social (including interaction between explainer and explainee), Truthful (all elements are true in the explanation with respect to the underlying system), and Thorough (describes the underlying system in its whole) using a 7-points Likert scale (1-not at all, 7-extremely). Since we purposefully designed the explanation for laypeople/non-expert, and not for domain experts, we associated a negative score to domain-dependent answers and formulated the question in a domain-independent manner. This means that if more users assessed the explanation as domain-independent, the explanation would be considered more meaningful.

To further examine the relation between meaningful explanation and trust, we asked participants to rate how much each explanation prototype component affects their understanding and trust towards the AI system. Specifically, participants were asked to rate to what extent each of the information provided in the explanation prototype is relevant for them to understand and trust the AI system, and to what extent each of the features provided by the prototype improved their understanding and trust in the AI system. The information and features provided in the prototype were a direct implementation of the Explanation Content and Explanation Interaction guidelines (EDGs), and were rated using a 5-points Likert scale. To assess if there were any changes in users’ perception and their trust judgements, we followed a within-subject design, in which we tested the difference between users’ trust levels before and after the interaction with the prototype.

To measure users’ trust, we used a measurement scale for human-AI trust in AI healthcare [54]. The scale has demonstrated good reliability ($\alpha > 0.88$) and is composed of six trust factor metrics: perceived understandability, perceived reliability, perceived technical competence, faith, personal attachment, and helpfulness. We asked participants to rate each of the metrics (Q3-8 and Q17-22). In addition to that, we also asked participants to rate their self-reported trust towards breast cancer self-assessment applications in general, before and after the interaction with the Care prototype (Q9 and Q23). Both the trust measurement scale and the self-reported trust question were rated in a 7-points Likert scale. Additionally, we asked the participants to fill out a System Usability Scale [13] for the prototype they had tested.

7.1.2 Second Survey. The second survey was conducted to refine and enhance the data collected in our study. While the first survey focused on the meaningful explanation characteristics and comparing participants’ perceptions before and after interacting with the explanation prototype, the second survey introduced an additional control group. This control group consisted of participants who did not receive any explanation in the prototype, allowing us to establish a baseline for comparison. By including the control group, we aimed to gain a deeper understanding of the impact of the
explanation prototype on user perceptions. This refinement in methodology enabled us to gather more comprehensive and nuanced insights into the effectiveness of the explanation prototype and its influence on user understanding and trust in comparison to the absence of an explanation.

The online survey containing three sets of questions and 40 items in total. The first set of questions covers three demographic items: gender, age, occupation; and two trust propensity questions. The second and third sets of questions each contain 16 questions from the Human-AI trust measurement instrument [55], and one question on subjective trust level. The instrument is the more refined version of the previous trust measurement. The instrument comprised of eight trust factors: perceived understandability, perceived reliability, perceived technical competence, faith, personal attachment, perceived helpfulness, institution credibility, and user autonomy, with two statement items for each factor (total 16 items). The reliability and validity of this measurement instrument were established. Between the second and third sets of questions, each participant was assigned to interact with one of the prototypes.

7.2 Data Collection and Analysis

7.2.1 First Survey. The participants were recruited on Mechanical Turk, with a survey set up using Google Form. We added a check-in question in the survey to maximise participation quality and avoid bot submission by checking if the participant read the text in the prototype carefully. Among the 82 subjects who took part, 22 failed to answer the check-in question, and there were also six duplicate entries. In total, 55 participants’ answers were analysed with a Mann-Whitney U test to investigate the change in trust for each metric. We then performed Bonferroni correction to control the false discovery rate (FDR).

7.2.2 Second Survey. Participants were recruited from Mechanical Turk and the survey was set up with Google Form. We chose "experienced worker" as the criterion and included a follow-up question within the survey to check whether participants were interacting and paying attention to the prototype properly. We targeted 50 participants in each group and the tasks lasted six weeks. After eliminating all duplicate and low-quality items, we had 88 participants in total. Participants’ answers were analysed with a Mann-Whitney U test to investigate the change in trust for each metric. Lastly, Bonferroni correction was performed to control the false discovery rate (FDR).

Each participant received a lump sum payment of 5 or 2 USD for prototypes with and without explanation, respectively. The study was approved by our institution’s Human Research Ethics Committee (HREC/4157).

7.3 Results

7.3.1 First Survey. We assessed to what extent the explanation provided by the prototype was considered as meaningful explanation to non-experts. More than half of the participants expressed their agreement that the explanation given possessed meaningful explanation characteristics (See Table 5). Participants highly rated the following characteristics: Contrastive, Domain-independent, General, Truthful and Thorough, with the Median = 6. Participants’ perception seems to be spread out with regard to the Social characteristic of explanation, with 50% (N=28) of participants unsure or slightly unsure if the explanation prototype contains interaction between explainer and explainee (rated 3-5).

<table>
<thead>
<tr>
<th>contrastive</th>
<th>domain independent</th>
<th>general</th>
<th>social</th>
<th>truthful</th>
<th>thorough</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>5.76</td>
<td>5.54</td>
<td>5.83</td>
<td>4.90</td>
<td>5.8</td>
</tr>
<tr>
<td>MEDIAN</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 5. Participant’s rating of explanation based on the characteristic of meaningful explanation.
As mentioned previously, to evaluate the prototype and see if it affects users’ trust judgment on the AI medical support system, we used a measurement scale for human-AI trust in AI healthcare composed of six trust metrics: perceived understandability, perceived reliability, perceived technical competence, faith, personal attachment, and helpfulness [54]. We then also asked participants to rate their self-reported trust towards breast cancer self-assessment applications in general, before and after the interaction with the Care prototype. Both the trust measurement scale and the self-reported trust were rated in a 7-points Likert scale.

We carried out a within-subjects study, and ran a Mann-Whitney U test to see any significant effects after the participants interacted with the prototype. Significant differences were observed in four out of six trust metrics at p < .05. There was a significant effect on perceived reliability of the AI system (Z-Score = 2.615, p-value = .0088); perceived technical competence (Z-Score = 3.183, p-value = .00148); personal attachment (Z-Score = -2.412, p-value = .015 and helpfulness (Z-Score = 2.188, p-value = .0285). Concerning the other trust metrics, two of them: perceived understandability (Z-Score = 2.188, p-value = .0285) and faith (Z-Score = 2.188, p-value = .0285), showed no significant differences.

In light of the multiple comparisons conducted in our analysis (six in total), we applied the Bonferroni correction to account for the increased risk of Type I errors. With the conventional significance level set at \( \alpha = 0.05 \), the corrected significance threshold would be \( \alpha_{\text{corrected}} = 0.05 / 6 = 0.0083 \). Upon applying the Bonferroni correction, we found that only one trust metric remained statistically significant. Specifically, perceived technical competence continued to exhibit a significant difference (p-value = 0.00148).

To illustrate the difference between users’ perception on six trust factors before and after the prototype interaction please see Figure 13. It can be seen that median values of each trust factor rating were moved towards the middle, with the exception of personal attachment, indicating the decrease in frequency on both ends. Additionally, the Care prototype reached an average score of 81.34 on the System Usability Scale, which falls in the "excellent usability" range [13].

From the analysis of the self-reported trust before and after interaction with the prototype we found that participants’ trust levels were not significantly different (Z-Score = -1.035, p-value = .298). Similar with the trend shown in trust factor rating, the trust level rating were also moved towards the middle after interaction with the prototype, with a decrease in frequency for trust level on distrust (rate = 1) and fully trust (rate = 7), as shown in Figure 14.

To evaluate the explanation prototype components, we asked non-experts to rate to what extent each information provided is relevant to understand and trust the AI system (1-not at all relevant, 5-very relevant). Most participants rated all of the information provided as "slightly relevant" or "very relevant" (Median=4 or 5) to understand and trust the AI system. In conjunction with the information provided in the explanation, we then asked participants to rate the improvement in their understanding and trust judgement of the AI system contributed by each feature of the explanation prototype. Using a 5-point Likert scale (1-very poor improvement, 5-very good improvement), all of the features were considered as "good improvement" (Median=4) to participants’ understanding and trust towards the AI system.

7.3.2 Second Survey. As mentioned in the method section previously, we added control prototype of no-explanation to answer the research question with between-subject design since the previous survey only compared within-subject groups. The prototype was designed as similar as the prototype with explanation which was developed based on EDGs (See Table 7).
We carried out a within-subjects study, and ran a Mann-Whitney U test to see any significant effects after the participants interacted with explanation prototype and after interaction with control prototype. We found no significant differences were observed in eight trust factors at $p < .05$. However, recognising the need to account for multiple comparisons and maintain the integrity of our analysis, we applied the Bonferroni correction. With six comparisons conducted in this phase of the study, the corrected significance threshold was set at $\alpha_{corrected} = 0.05 / 6 = 0.0083$. 

Table 6. Participant’s rating of information and features in the explanation prototype relevancy to understand and trust the AI system.

<table>
<thead>
<tr>
<th></th>
<th>disease information</th>
<th>treatment</th>
<th>next plan/step</th>
<th>data info</th>
<th>input</th>
<th>system process</th>
<th>output</th>
<th>comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>median</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>mean</td>
<td>4.65</td>
<td>4.43</td>
<td>4.54</td>
<td>4.10</td>
<td>4.21</td>
<td>4.38</td>
<td>4.67</td>
<td>4.38</td>
</tr>
<tr>
<td></td>
<td>doctor appointment</td>
<td>input check</td>
<td>detail request</td>
<td>open question</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>median</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>4.07</td>
<td>3.83</td>
<td>4.21</td>
<td>4.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 13. Participant’s rating of trust factors before and after interacting with the prototype.
Meaningful Explanation Effect on User’s Trust in an AI Medical System

Fig. 14. Participant’s rating of trust before and after interacting with Care prototype.

<table>
<thead>
<tr>
<th>Explanation Design Guidelines (EDG)</th>
<th>Explanation Prototype</th>
<th>Control Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDG1: Disease Information</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EDG2: Disease Treatment</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG3: Next Plan/Step</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG4: System Information</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG5: System Input</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EDG6: System Process</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG7: System Output</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EDG8: Empathy (Reassuring Words)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EDG9: Simple and General</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EDG10: Input Check</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG11: Doctor Appointment</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG12: Open Question</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG13: Input Comparison (Visualisation)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>EDG14: Detail Request</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 7. Explanation Designs Guidelines Implementation on Prototype with and without explanation (control).

...we found that none of the trust factors exhibited a significant difference between participants’ interactions with the explanation prototype and the control prototype. This suggests that, even after correction, there were no statistically significant variations in trust factors between the two experimental conditions.

Furthermore, we aimed to assess the individual-level impact of explanations on users’ initial trust. To achieve this, we conducted a comparative analysis within the same group, examining the effect of explanations before and after users interacted with the prototypes, similar to the approach employed in the First Survey. The Mann-Whitney U test was utilized to determine potential differences in trust levels, with separate tests conducted for each trust factor. Interestingly, no significant differences were observed between trust ratings before and after the interaction. As illustrated in Table 8,
the data suggests that the provision of explanations did not lead to substantial changes in users’ initial levels of trust.
These findings highlight the inherent limitations of explanations in altering users’ initial trust perceptions, emphasizing
the need to explore alternative strategies to address this aspect of user trust.

8 DISCUSSION AND REFLECTION

8.1 Findings: Guidelines and Effect on Trust

In this paper we propose a set of explanation design guidelines for non-expert users of AI medical systems. With
the involvement of medical domain experts, AI experts and non-experts in several stages of design, our explanation
models and guidelines contribute a valuable new perspective to the human-centered explainable AI research [1, 4].
More specifically, our interpretation of explanation design in the light of diverse human expertise and targeted to
non-experts, adds to other Human-AI guidelines, which built on experts’ or practitioners’ experience only [2, 77, 78, 91].

Far from being an ultimate guidelines list, this can be improved by incorporating other AI design guidelines currently
emerging in the AI literature. For example, Amershi et al.’s guidelines [7] for human-AI interaction indicated that AI
should show contextually relevant information (G4) and mitigate social biases (G6). These guidelines could be added to
extend our guideline EDG9: Simple and General, to put more consideration in writing general explanations which are
aware of and mitigate social biases. Even though some of the current AI explanation guidelines are too general and
claimed to be not actionable [57, 91], a critical literature review would help to expand and refine our explanation design
guidelines.

We also discovered that the explanation provided by the prototype covered most of the theory-based characteristics
of meaningful explanations. An interesting finding of our study concerns the non-expert users’ ratings of the domain-
dependent and social characteristics of explanation. According to the literature, people consider an explanation to be
understandable if they can find a connection with their own domain knowledge or role, which implies that a meaningful
explanation is domain/role dependent [69]. However, our results suggest the opposite. Participants mostly agreed that
they can understand an explanation without requiring additional knowledge. This rather contradictory result may
therefore suggest that the domain/role dependent characteristic of understandable explanation is correct only to a
certain extent. For example, it clearly is the case that an explanation from an AI medical support system would be
perceived differently by a layperson compared to a medical expert. However, if the comparison is made between two
laypeople who have different background knowledge unrelated to the medical field, they may both experience the same
understanding. This result could also be caused by how the explanation is designed. As described in the previous section
(Section 5.2), the explanation was intentionally written to follow EDG9: Simple and General, where we minimised the
number of medical and technical terms used in it.

Another interesting result concerns the characteristic of explanation as a social exchange that involves interaction
between explainer and explaine [38, 73]. We expected the result to be in between the 1-3 range, because the prototype
we developed is quite static and did not include a dialogue feature or conversation to simulate social exchanges. However, the rating was higher than we expected (median=5). This result might suggest that the current level of interactivity is considered as an adequate social exchange for user. However, there is also the possibility that the participants did not really understand the question, or may not have had examples of more interactive systems in the past. Further investigation of this aspect and a possible rephrasing of the question might be considered. Nonetheless, overall, the positive sentiment both in the qualitative and quantitative evaluation indicate the effectiveness of the prototype in presenting a meaningful explanation to non-expert users.

We also investigated the relation between meaningful explanation and trust. Our evaluation showed that exposure to the explanation prototype moved non-expert users’ trust rating towards the middle range for the overall trust, along with the extremes rating on both ends (distrust and fully trust) reducing in numbers (See 14). With the distrust and fully trust frequency decreasing, our hypothesis that meaningful explanation can help users make more considered trust judgements (and not over-trust or fully distrust an AI system) is supported.

However, we recognise that trust is difficult to measure and there are broad definitions of trust. A simple self-report trust scale might not seem enough to be used, and previous studies claimed that self-reported trust is not a reliable measure for trusting behaviors [52, 89, 105]. These claims highlight the focus area of this research, where we have specifically looked at trust as an attitude rather than trust related behaviour [18, 88]. This distinction is important, since trust attitude does not always correlate to trust related behavior [65, 72].

The broad definition of trust also resulted in a different trust measurement method. Previous research on explanation for non-experts [21, 46, 81, 110] have shown mixed results in terms of users’ trust. We believe that there are possible factors, such as different evaluation methods, the lack of universal measurement metric, and unclear definitions of trust, which may have affected those different results. According to Vereschak et al. [101], research on trust should provides a clear definition of trust and specify the terminology used to prevent any confusion between trust and trust related constructs, such as confidence, reliance, compliance and distrust.

Other than trust as an attitude, we also measured trust related constructs: factors that affect human-AI trust. The scale measures perceived understandability, perceived reliability, perceived technical competence, faith, personal attachment, and helpfulness [54], and the analysis showed that our designed explanation prototype affected these factors. From the first survey, four out of six trust factor ratings were significantly different. Even though the effect of meaningful explanations was not significant on the second survey, this does not diminish the importance of providing explanations to help the user calibrate their level of trust. Both on first and second survey, for the overall trust level, the extreme ratings on both ends were reduced (See 13). This result further suggest the potential of meaningful explanations to help users consider their trust judgements. However, this result need to be taken with caution, since participants rating and perception are surely affected not only by the explanation but also by the prototype UI and UX (how the explanation is communicated, delivered and customised).

The guidelines we propose are theory based, multi users-validated and pragmatically applied. We implemented the 14 guidelines in a prototype, carried out a mixed method evaluation of the prototype. Evaluation results suggest that the explanations provided by our prototype were meaningful and could effectively moderate trust judgements toward the AI system (by reducing higher trust and improving lower trust judgements). This evidence the effectiveness of our prototype design and provide an initial validation of our explanation design guidelines for non-expert users of AI medical systems. The biggest potential of our explanation design guidelines implementation is, the closest scenario,
for self-managed breast cancer assessment apps, already available on the market\textsuperscript{8,9,10}, or for post-hoc explanation in various deep learning approaches \cite{87}. This is not to close implementations in other scenarios, such as, skin cancer detection \cite{48, 95} or mental-health care \cite{17, 42}.

From a methodological perspective, we successfully conducted a stage-based design process which involved eliciting different stakeholders’ point-of-views and distilled them into explanation models and design requirements. In the explanation model development phase, we found that some explanation components that were present in the expert explanation model are not present in the user explanation model, and vice versa. However, at the target explanation model stage experts’ and non-experts’ opinions were reconciled. Later in the evaluation stage, there were no significant disagreement in the explanation components (what to explain) and in the explanation presentation (how to explain). This shows that the method is quite successful in getting diverse ideas together. The main structure of the original method is also adaptable, following different explanation goals and different models. This method is one of the few explainable AI’s requirement elicitation methods \cite{36, 47, 57}, which are versatile to different explanation design goals. However, the method is not without faults. The original method we adapted is a stage-based participatory design process, and participatory design processes are known to take extensive amount of time and resources \cite{94}. The research summarised in this paper required two years of intense design, implementation and various cycles of multi method evaluations (several interviews analysed with QDA methods, both inductive and deductive, and quantitative analysis of online survey data). The whole participatory design method consisted of five stages, and each stage needed extensive time for research design, data collection and analysis, which takes several months to complete.

8.2 Limitation and Future Work

There are a number of limitations that are important to mention from each research stage described in the paper. In the explanation models development, experts and non-experts recruited were from the researchers’ personal and social network, which might affect the diversity of views. We foresee the value of future work to involve a bigger number of domain experts and AI experts with different and broader levels of expertise. There is also a possibility of researcher bias when implementing grounded theory approach in the analysis, since the model development stages were done in sequence, and the expert explanation model analysis could affect the analysis for the user explanation model.

In the guidelines development and implementation stage, we recognise that the prototype development only involved designers, without any user input, which might have caused bias. Further research is necessary to understand how user input might affect the prototype and how other designers might employ these guidelines for different AI healthcare cases. Additionally, we did not include iterative cycles between prototyping and evaluation. Iterative cycles of implementation and evaluation could be potentially valuable to gather more refined results of explanation and prototyping. Implementations of different AI system applications, which use our proposed explanation design guidelines are also required to obtain a more robust understanding of how different types of explanations impact the user’s ability to make considered trust judgements.

In the evaluation stage, as mentioned in the discussion section previously, we only evaluated users’ trust as an attitude, not including trust related behaviour. Future studies with additional trust measurement, such as, decision to trust, act of trusting, perceived trustworthiness, and trust related behaviour will be beneficial to form a complete view on the effect of meaningful explanation on trust \cite{18, 18, 65, 72}. For example, in the case of a self-managed breast

\textsuperscript{8}https://talosapp.me/breast-cancer-self-exam-self-screening-talos-app/

\textsuperscript{9}https://www.braster.eu/en/system-braster/what-is-braster

\textsuperscript{10}https://www.niramai.com/
cancer related AI system, we should investigate if whether a layperson would be confident and follow through the AI system’s recommendation, and whether they would fully rely on the AI system’s diagnosis or recommendation. We also evaluated users’ trust using trust measurement based on six trust factors. We recognised that this measurement is not complete, and additional factors that could affect trust in human-AI interaction might be needed [33, 70]. This research also only followed a within-subject study and did not include a control prototype. A study with baseline explanation or control prototype and between-subject analysis would provide more insight about the effect of meaningful explanation.

The evaluation data was collected using MTurk. We understand that there is possible inattention, self-misrepresentation, self-selection bias, and social desirability bias inherited in the data, which may affect the overall result [5]. Additional qualitative data collection with the same questions could minimise the risk and add more depth to the evaluation. We also used a fake third-person profile and a hypothetical tested scenario in the prototype. Since this scenario lacked the significance of a real-world decision, this could affect how participants answered the questions. Future research carried out in a real clinical trial would be tremendously insightful to validate our findings. Finally, the mixed-method evaluation consisted of one online survey and seven semi-structure interviews. Future studies with a bigger sample size are needed to ensure the reliability of the results.

9 CONCLUSION

In the context of user-centered approaches to AI explanation, we provide several valuable contributions to inform future research in the field. We adapted a five stages participatory design process to distill key characteristics of AI explanations that are adequate and understandable, which we defined as meaningful explanations. We developed three explanation models: Expert Explanation Model, User Explanation Model, and Target Explanation Model. We translated the Target Explanation Model into 14 design guidelines for meaningful AI explanations. The explanation guidelines were then used to inform the design of an AI explanation system prototype for non-expert users in a breast cancer self-managed health scenario. In the evaluation of the prototype, we found out that non-expert users’ trust levels and trust factors perceptions were changed after interacting with the AI explanation prototype. These findings suggest that meaningful explanations can help non-expert users to make more considered judgements on trusting an AI system in a healthcare scenario (in particular by moderating users’ trust levels). Finally we contributed reflections on the limitations of our study that could be further explored and inform new opportunities for researchers in the field to study and realise trustworthy explanations for AI healthcare systems.

REFERENCES


1872

1871

1870

1869

1867

1866

1865

1864

1863

1862

1861

1860

1859

1858

1857

1856

1855

1853

1852

1851

1850

1849

1848

1847

1845

1841

1840

1839

1838

1837

1835

1834

1833

1829

1828

1827

1826

1825

1822

1821

13 Larasati, et al. Manuscript submitted to ACM


Meaningful Explanation Effect on User’s Trust in an AI Medical System


Manuscript submitted to ACM
Meaningful Explanation Effect on User’s Trust in an AI Medical System


